Scott Schmidt; Data Science; 10/4/2021

**1. How does this position fit with the career path you envision for yourself?**

I believe Python is a superpower, and I want to help people by making data driven decisions.

The ISU data science team will allow me to grow as a programmer and learn from my peers as I continue to use the Python and data science skills gained from prestigious hedge funds to benefit the education system. The first step of a data scientist is knowing and being able to understand the data.

Illinois State University is home to me as an alum, employee, former graduate assistant, current student, and future student, and I believe the data science team will give me the opportunity to apply my knowledge as someone who knows the university culture.

**2. Data extrapolation and analysis are critical to accurate models being developed. Please describe an example of data extract and analysis that you have completed.**

Dropping null values, outliers, and data cleaning is a large portion of a data science role.

At Illinois State University, I invented a Python data science pandas dataFrame automation program that

compares thousands of data transactions to database records which detects unbalanced accounts, new transactions, and human error:

<https://github.com/ScottFrederickSchmidt/AccountingAutomation>

This project involved a hundred lines of code with most of it involving data cleaning. For instance, the hyphens in the real data had to be temporarily dropped to match the Access database records. Accounting records had to be grouped together. This was an important task because it involves accounting. If the program is not 100% accurate, one could get involved in an audit that could cost the company thousands of dollars and a bad reputation. In accounting, one must be exact. There are no guesses. Accuracy, attention to detail and perfection is of highest importance. My accounting software can verify if an account is not balanced with 100% accuracy. If my accounting software was implemented on a global basis, I think it would save ISU over one hundred thousand dollars a year. It would also not automate jobs but instead make the job of an accountant more enjoyable.

Likewise, at Douglas Capital Management I programmed Python financial automation using pandas, ib\_insync, and operating systems (os). One slight mistake with the trading bot could mean a loss of money if the bot gathered incorrect information using the ib\_insync package. I also used Python to webpage data scrapping using BeautifulSoup (bs4), requests, selenium, webdriver, and to\_csv. For example, I developed a weather bot to help get the average temperature across major cities in the United States to help determine the fair value of the price for natural gas depending on the weather during winter. Also, one must make sure that the data is correct. For instance, a stock split could easily distort data if the data does not do adjustments. Many times, investment firms (and other companies)

purchase data that is bad data without realizing it. Another example of understanding data would be if a job title changed titles ten years ago. To gather the most data, one would have to know this information to use that previous job title data. Otherwise, the data will not be using lots of datapoints that could have been collected in the predictive model.

Coding efficiency, queries, and data manipulation is my strength as I have over one thousand hours in Python data manipulation which can be viewed on both GitHub and YouTube:

<https://github.com/ScottFrederickSchmidt/Project-Euler-Python>

<https://github.com/ScottFrederickSchmidt/LeetCode>

For instance, my first solutions to problem 44 took 86 seconds. After putting many hours of changing the logic (i.e. when for a greater number is searched and adding in a function), I was able to get the code to run in under two seconds with a new, revised solution.

**3. Please describe the experience you have creating predictive models.**

Predicting stock market returns was the entire goal at Douglas Capital Management. I used portfolio123.com database with hundreds of finance formulas to predict future returns. For instance, OperCashFlTTM - CapExTTM + IntExpTTM\*(1-TaxRate%TTMInd/100) would be inserted into a database with an optimization process to see how this one formula can affect future returns. More important formulas such as annual growth rate would have a higher significance for returns. We also had to account for missing data. On occasion, formulas had to be completely deleted if there were too many missing values. But most of the time, missing data was filled in using an estimate by a regression line. Basic MATLAB functions would be used to help find the statistic relevance of each formula. In 2019, our returns were in the top ten. During this time, I also designed Python financial automation using pandas and os that would teach an AI bot when to place a trade.

During my MBA program as an ISU graduate assistant, I generated a 95% r-squared using data analytics (depends on adjustment to overfitting). This identified the best candidates that would most likely successfully graduate at ISU. This was the final MBA capstone course, MQM497. In conclusion, Pell Grants and number of initiated contacts are the two biggest factors in whether a student enrolled at the university or not. These two factors alone can make a good prediction on whether a student enrolls at the university or not. In conclusion, neural networks show that undecided majors, Pell Grant, distance from ISU, # of initiated events, family income median, total consumer expenditures, high school low-income percentage, and high school GPA are the major factors in whether a student enrolls at Illinois State University or not. The test confirms that ACT, GPA, major, Pell Grant, and # of initiated events make a significant difference in students enrolling at Illinois State University.

Creating a prediction starts with removing missing values and outliers. Data cleaning is also a timely but important process. At Clark Street Capital, we used pandas, numpy, and sklearn to predict loan credit risk. Credit risk was predictive through variables such as employment, zip code and credit score. This risk was important to analyze because loans were bundled together by their risk score.

My "Amazon" website is built on SQL and has a SQL recommendation system that would predict what purchases the customer might want to buy next:

<https://github.com/ScottFrederickSchmidt/Amazon-like-site>

Sample predictive Python projects using numpy, pandas, matplotlib.plplot, and sklearn.model\_selection can be found my GitHub below:

<https://github.com/ScottFrederickSchmidt/PredictivePythonModeling>

**4. In terms of SQL syntax or query structure, what are some things you focus on when writing a query to ensure it functions as efficiently as possible?**

To write an effective SQL query, one first needs to understand the data (hopefully I have an advantage at understanding Illinois State University data being employed herefor almost five years now). Once the data is understood, the criteria must then be identified. Then, one can now create an efficient SQL query to find the desired results. Here are some ideas of producing an efficient SQL result:

The larger the table, the longer it takes to read the data. Using a JOIN statement will increase the runtime. As such, I am careful to use only the tables needed. One should try to only use the columns needed. Using inner join can limit the rows brought into the data to only what is needed. Using subqueries should be used whenever possible before using a JOIN statement if the subquery is returning a single value. One should never use the \* operator. Including specific column names enhances performance. It also makes the code more reusable in the future (if more columns get added later). Using the \* function is also a security issue. SQL can accidentally gather confidential information such as addresses, studentID, age, birthday, and confidential information into an csv file when it should not be there. One should not repeat code. In my automation project, I made sure to never use a line twice. Temporary tables should be used to speed up efficiency instead of having multiple subqueries referring to the same table.

On my "Amazon" site, I built an entire SQL database that was on the backend. It is connected on the frontend using PHP. The basic SQL terminology were all used for this project (i.e. select, insert, update, delete). In addition, it has a basic SQL search and recommendation section. All these SQL queries happened with high performance speed: <https://github.com/ScottFrederickSchmidt/Amazon-like-site>

Memory is an important issue of the life of a programmer. This is because Excel has a max of around two million rows while Python can have a max out on memory with very large, complex dataFrame. However, having rows at this magnitude does not happen often. This is when SQL data manipulation needs to be done, which is generally more challenging. Challenging SQL analytics was done in my Udemy course which can be found below:

<https://github.com/ScottFrederickSchmidt/Complete-SQL-Bootcamp-and-SQL-DataAnalytics-Course>

Coding efficiency, queries, and data manipulation is my strength as I have over one thousand hours in Python/SQL data manipulation, but mostly in Python. This can be viewed on both GitHub and YouTube below:

<https://github.com/ScottFrederickSchmidt/Project-Euler-Python>

<https://github.com/ScottFrederickSchmidt/LeetCode>

**5. Part of the Data Scientist role will be to provide visualizations that will be able to explain the analysis that you have performed. What tools have you used to create data visualizations?**

I have mostly used Python, Excel, PowerPoint and Tableau for visualizations. In a recent data science

poll on LinkedIn I saw Excel was ranked as being "highly underrated" for data science.

Tableau gives a data scientist an easy way to explain data to an individual who may not have a statistics or coding background. In fact, Tableau makes it easy for even developers to understand data and make analytical decisions.

Because I have a web development background, I could also display graphs on the Illinois State University website if needed with a JavaScript API such

as https://www.chartjs.org/docs.

**6. Please describe a presentation that you have completed to present the results of a predictive model you developed.**

At Clark Street Capital, I had to help package and sell at $42 million-dollar non-performing commercial real estate loan bundled together as a mortgage-back security through a major United States bank. Based upon our predictive analysis of the loan risk, the goal was to sell the loan around eighty cents on the dollar. I had to help present this MBS product over phone to over a hundred banking clients across

the United States.

At Illinois State University, I invented a Python data science pandas dataFrame automation program that

compares thousands of data transactions to database records which detects unbalanced accounts, new transactions, and human error:

<https://github.com/ScottFrederickSchmidt/AccountingAutomation>

I have presented this project to several programming professors and directors because I have wanted to try to implement my program on a more frequent and global basis. I truly believe my program can save ISU over a hundred thousand dollars a year if correctly implemented. It will also make the life of an accountant more enjoyable. Presenting programming code is more of a challenge than I originally thought because I have concluded that data manipulation and programming are almost two completely different specialized subjects.

Sample predictive Python projects using numpy, pandas, matplotlib.pyplot, and sklearn.model\_selection can be found my GitHub below:

<https://github.com/ScottFrederickSchmidt/PredictivePythonModeling>

**7. Have you ever had to “sell” an idea to your co-workers, supervisor or customers? How did you do it?**

Great question! I think my data science accounting automation software I have developed would save Illinois State University, if implemented across the entire university, well over a hundred thousand dollars a year. I am currently writing a proposal that will be send to the ISU administration on my project.

A prepared prototype using fake data was built before hand so I had confidence the program would work. When I did invent my automation program, I did request permission to use the data first.

I also made sure that no confidential information would accidently be leaked.

**8. During analysis, how do you treat missing values?**

The simple, easy way to treat missing values is to simply delete any data row that has a missing value using a simple Python script such as df.dropna(). This is my preferred method. This method does not make assumptions on top of assumptions.

Another way is to find the mean of the data and assume that the missing value is that number. KNN nearest algorithm can be used to impute data.

A bad choice would be to fill the missing data with a 0 because this would distort the data.

Lastly, if there is not a lot of data points with lots of missing data, one should consider not using that data at all. A lack of data points will not create an accurate prediction.

After writing out my answer, I realized that on July 2nd, 2020, I wrote a brief article on my GitHub on this exact question which can be viewed here:

<https://github.com/ScottFrederickSchmidt/PythonDataScience/blob/master/dropna()%20or%20fillna()>

**9. What are the tools you have most commonly used to complete predictive analysis? Which one is your favorite?**

I like using Python because it has a wide variety of tools such as pandas, numpy, and sklearn.

This is what I used to automate the accounting analytics predictions at Illinois State University,

financial automation at Douglas Capital Management, and predictive analytics at Clark Street Capital.

This is what was taught to me by my data science mentor, Sukumar Lagapati and used in my Python data science bootcamp:

<https://github.com/ScottFrederickSchmidt/PythonDataScience>